線形言語のある部分言語族に対する質問と特徴的なサンプルによる多項式時間学習アルゴリズム (計算機科学基礎理論の新展開)

AUTHOR(S):

但馬, 康宏; 小谷, 善行; 寺田, 松昭

CITATION:

但馬, 康宏 ...[et al]. 線形言語のある部分言語族に対する質問と特徴的なサンプルによる多項式時間学習アルゴリズム (計算機科学基礎理論の新展開). 数理解析研究所講究録 2004, 1375: 99-105

ISSUE DATE:

2004-05

URL:

http://hdl.handle.net/2433/25581

数理解析研究所講究録 1375 巻 2004 年 99-105

線形言語のある部分言語族に対する質問と特徴的なサンプルによる多項式時間学習アルゴリズム

但馬 康宏，小谷 善行，寺田 松昭
Yasuhiro TAJIMA, Yoshiyuki KOTANI and Matsuaki TERADA

東京農工大学 情報コミュニケーション工学科
Department of Computer, Information and Communication Sciences,
Tokyo University of Agriculture and Technology

1 はじめに

本研究において，線形言語のある部分言語族に対する，所属性質問と学習対象を特徴づける正の例集合 (代表部分集合) を用いた多項式時間厳密学習アルゴリズムを示す．

このアルゴリズムの時間計算量は，学習対象を表すことのできる文法のサイズ，与えられたサンプルの最大長，および学習対象を表すことのできる文法族の等価性判定に必要な時間に関する多項式である．

同様の条件のもとでは，単純決定性言語族に対する多項式時間学習可能性が示されている [6]．この単純決定性言語族に対する学習アルゴリズムは，複数の仮説文法を効率的な等価性判定を用いて絞り込む点が特徴的であるが，本研究におけるアルゴリズムは，その手法を線形言語の部分族に対して適用することにより得られる．本研究における学習対象の言語族は，正則言語族を真に含み，単純決定性言語族とは比較不能である．

線形言語の部分族に対する学習アルゴリズムについては，正則言語の学習に帰着させる手法の研究 [5] や学習達成条件の緩和に関する研究 [4] などが示されている．特に文献 [4] では，学習対象を特徴づける記号列集合を含んだサンプル集合を与えられた場合に正しい仮説を提示すればよいとしており，これは，本研究における学習可能性から直ちに導かれる．

本研究で学習対象としている言語族は，これらの関連研究で学習対象としている言語族を含むか比較不能な関係にある．

2 諸定義

文脈自由文法を $G=(N,\Sigma,P,S)$ で表す．ここで N は非終端記号の有限集合，Σ は終端記号の有限集合，P は生成規則の有限集合，$S\in N$ は開始記号である．ε を空記号列とし，$A\to\varepsilon$ なる生成規則が文法中に存在しない場合，ε-規則なしであると呼ぶ．G におけるすべての生成規則が $A\to a\beta$ なる形であるとき，G を Greibach 標準形と呼ぶ．ここで，$A\in N$, $a\in\Sigma$, $\beta\in N^*$ である．

文脈自由文法 $G=(N,\Sigma,P,S)$ は，以下の条件を満たすとき 2-標準形 [2] であるという．

- P は，Greibach 標準形である．
- すべての生成規則 $A\to a\beta$ は，$|\beta|\leq 2$ である．

このとき，任意の文脈自由言語は 2-標準形の文脈自由文法で表すことができる [2]．

生成規則の集合 P が ε-規則なしであり，さらにすべての規則が以下のいずれかの形の場合，G を線形文法と呼ぶ．(1) $A\to aBc$, (2) $A\to aB$, (3) $A\to Bc$, (4) $A\to c$. ここで，$A,B\in N$ であり，$a,b,c\in\Sigma$ である．

生成規則 $A\to a\beta\in P$ および $\gamma,\gamma'\in(N\cup\Sigma)^*$ に対して，$\gamma A\gamma'$ から $\gamma a\beta\gamma'$ への導出を $\gamma A\gamma'\underset{G}{\to}\gamma a\beta\gamma'$ と表す．$\alpha,\alpha'\in(N\cup\Sigma)^*$ に対して 0 回以上の導出を $\alpha\overset{*}{\underset{G}{\Rightarrow}}\alpha'$ と表し，文法 G が明らかなときは，単に $\alpha\overset{*}{\Rightarrow}\alpha'$ と表す．

任意の $\gamma\in(N\cup\Sigma)^*$ に対して $L_G(\gamma)=\{w\in\Sigma^*\mid\gamma\overset{*}{\underset{G}{\Rightarrow}}w\}$ を γ の生成する言語と定義する．特に $L_G(S)$ を G の生成する言語と定義し，$L(G)$ と表す．

文脈自由文法 $G=(N,\Sigma,P,S)$ において，ある $\alpha,\beta\in(N\cup\Sigma)^*$ に対して $S\overset{*}{\underset{G}{\Rightarrow}}\alpha A\beta$ なる導出が可能な $A\in N$ を到達可能な非終端記号と呼び，$L_G(A)\neq\emptyset$ なる $A\in N$ を live な非終端記号と呼ぶ．

線形文法 G および任意の $w\in L(G)$ に対して，w に対する導出が高々 1 通りしか存在しない場合，G を曖昧でない線形文法と呼び，そのような文法で表される言語を曖昧でない線形言語と呼ぶ．

その他の定義については，文献 [2][3] に準じるものとする．

定義 1 $G=(N,\Sigma,P,S)$ は，曖昧でない線形文法であるとする．$Q\subseteq L(G)$ は，以下の条件を満たすとき G の代表部分集合 (representative sample) と呼ばれる．

- G のすべての生成規則 $A\to\beta$ それぞれについて，$S\overset{*}{\Rightarrow}\gamma_1A\gamma_2\underset{G}{\to}\gamma_1\beta\gamma_2\overset{*}{\Rightarrow}w$ を満たすある $w\in Q$ が存在する．ここで，$\beta,\gamma_1,\gamma_2\in(N\cup\Sigma)^*$ であり，$A\in N$ であるとする． □

ここで G が曖昧でない線形文法であるとき，任意の $w\in L(G)$ に対してその導出木が一意に定まるので，代表部分集合は “すべての生成規則を使用しなければ生成できないような $L(G)$ の部分集合” である．

また，曖昧でない線形言語に対する代表部分集合を以下のように定義する．

定義 2 曖昧でない線形言語 L について，$Q\subseteq L$ が代表部分集合であるとは，Q が $L(G)=L$ かつ曖昧でないある線形文法 G の代表部分集合であるときとする． □

曖昧でない線形文法の代表部分集合については，以下の定理が成り立つ．

定理 3 曖昧でない線形文法 $G=(N,\Sigma,P,S)$ の代表部分集合は，$|N|$ および $|\Sigma|$ の多項式時間で構成可能である．

(証明) すべての $A\in N$ について，以下のような終端記号列 $u_A,y_A,w_A\in\Sigma^*$ をひとつずつ定める．

- $S\overset{*}{\Rightarrow}u_AAy_A$ であり，かつ $S\overset{*}{\Rightarrow}uAy$ であるようないかなる $u,y\in\Sigma^*$ に対しても $|u_A|+|y_A|\leq|u|+|y|$ である．
- $w_A\in L_G(A)$ であり，かつ $w\in L_G(A)$ なるいかなる $w\in\Sigma^+$ に対しても $|w_A|\leq|w|$ である．

代表部分集合 Q を以下のように定める．

$$\begin{aligned}Q\ =\ &\{u_Aay_A\mid A\in N,A\to a\text{が }P\text{に含まれる}\}\\&\cup\{u_Av_Bw_Bx_By_A\mid A\in N,v_B,x_b\in\Sigma^*,\\&A\to v_BBx_B\text{が }P\text{に含まれる}\}\end{aligned}$$

□

以下のような制限のある線形文法 $G=(N,\Sigma,P,S)$ を考える．

1. ある $A,B\in N$ および $a,c\in\Sigma$ について，$A\to aBc$ なる生成規則が P に含まれるならば，$B\neq C$ であるいかなる $C\in N$ についても，$A\to aCc$ もしくは $A\to aC$ なる形の生成規則は P に存在しない．
2. ある $A,B\in N$ および $a\in\Sigma$ について，$A\to aB$ なる生成規則が P に含まれるならば，以下の条件がすべて満たされている．
   - $B\neq C$ であるいかなる $C\in N$ および，すべての $c\in\Sigma$ について，$A\to aC$ もしくは $A\to aCc$ なる形の生成規則は P に存在しない．
   - さらに，いかなる $D\in N$ および，いかなる $b\in\Sigma$ についても，$A\to Db$ なる形の生成規則は P に存在しない．
3. ある $A,B\in N$ および $a\in\Sigma$ について，$A\to Ba$ なる生成規則が P に含まれるならば，以下の条件がすべて満たされている．
   - $B\neq C$ であるいかなる $C\in N$ および，すべての $c\in\Sigma$ について，$A\to Ca$ もしくは $A\to cCa$ なる形の生成規則は P に存在しない．
   - さらに，いかなる $D\in N$ および，いかなる $b\in\Sigma$ についても，$A\to bD$ なる形の生成規則は P に存在しない．

このような制限のある線形文法の部分族を偏導出線形文法族と呼び，この文法族で生成される言語族を偏導出線形言語族と呼ぶ．このとき，以下の性質を満たす．

定理 4 偏導出線形言語族は，正則言語族を真に含む．

(証明) 正則文法 $G_r=(N_r,\Sigma,P_r,S_r)$ の生成規則は，

すべて $A \to aB$ なる形か，$A \to a$ なる形である．ここで，$A, B \in N_r$, $a \in \Sigma$ とする．さらに，$A \to aB$ が P_r に含まれれば，$B \neq C$ であるいかなる $C \in N_r$ に対しても，$A \to aC$ なる生成規則は P_r に含まれない．したがって，正則文法の生成規則は，偏導出線形文法の制限の条件をすべて満たすので，すべての正則文法は偏導出線形文法である． □

定理 5 偏導出線形言語族は，単純決定性言語族と比較不能である．

(証明) 単純決定性言語と線形言語は，比較不能であるので，単純決定性文法で表せて，線形文法で表せない言語が存在する．さらに，言語 $L_1 = \{a^i b^i \mid i \geq 1\} \cup \{a^i c^i \mid i \geq 1\}$ は単純決定性文法で表せず，偏導出線形文法で表せる． □

定理 6 偏導出線形文法を $G = (N, \Sigma, P, S)$ とし，$w \in L(G)$ とする．ここで，$w_1 a u b w_2 = w$ であるような，$w_1, w_2 \in \Sigma^*$, $a, b \in \Sigma$, $u \in \Sigma^+$ について，

$$S \overset{*}{\underset{G}{\Rightarrow}} w_1 A w_2$$

まで導出を行ったとする．このとき，A から次の導出に適用可能な生成規則は，P の中から一意に決定可能である．

(証明) 偏導出線形文法の定義から，$A \to aB$, $A \to Bb$, $A \to aBb$ なる形の生成規則のうち，高々 1 つしか P に存在しない．したがって，題意を満たす． □

以後，学習対象の言語を L_t とし，$G_t = (N_t, \Sigma, P_t, S_t)$ を $L_t = L(G_t)$ なるある偏導出線形文法とする．また，$L_t \neq \emptyset$ であるとする．

学習対象の言語 L_t に対する所属性質問 $MEMBER(w)$ を以下のように定義する．

入力：任意の終端記号列 w

出力：yes ... if $w \in L_t$
　　no ... if $w \notin L_t$

3 学習アルゴリズム

3.1 学習アルゴリズムの基本戦略

学習者は，学習対象 L_t に対する所属性質問を行え，かつ学習開始時に代表部分集合 Q を得られるものとする．本研究における学習アルゴリズムは，単純決定性言語族に対する所属性質問と代表部分集合による学習アルゴリズム [6] と同様の手法により構成される．

ある終端記号列 $w \in \Sigma^+$ に対して，文法 $G_w = (R_w, \Sigma, P_w, S_w)$ を考える．ここで，$R_w = \{(u_1, u_2, u_3) \mid u_1, u_3 \in \Sigma^*, u_2 \in \Sigma^+, u_1 u_2 u_3 = w\}$, $S_w = (\varepsilon, w, \varepsilon)$ であり，

$$\begin{aligned}
P_w \;=\; & \{(v_1, a, v_3) \to a \mid a \in \Sigma, v_1, v_3 \in \Sigma^*, \\
& \quad (v_1, a, v_3) \in R_w\} \\
\cup\; & \{(v_1, a v_2, v_3) \to a \cdot (v_1 a, v_2, v_3) \mid a \in \Sigma, \\
& \quad v_1, v_3 \in \Sigma^*, v_2 \in \Sigma^+, (v_1, a v_2, v_3), \\
& \quad (v_1 a, v_2, v_3) \in R_w\} \\
\cup\; & \{(v_1, v_2 a, v_3) \to (v_1, v_2, a v_3) \cdot a \mid a \in \Sigma, \\
& \quad v_1, v_3 \in \Sigma^*, v_2 \in \Sigma^+, (v_1, v_2 a, v_3), \\
& \quad (v_1, v_2, a v_3) \in R_w\} \\
\cup\; & \{(v_1, a v_2 b, v_3) \to a \cdot (v_1 a, v_2, b v_3) \cdot b \mid \\
& \quad a, b \in \Sigma, v_1, v_3 \in \Sigma^*, v_2 \in \Sigma^+, \\
& \quad (v_1, a v_2 b, v_3), (v_1 a, v_2, b v_3) \in R_w\}
\end{aligned}$$

である．

文脈自由文法 $G = (N, \Sigma, P, S)$ において，導出木の非終端記号のラベルをすべて 1 種類の特別な記号 $\delta \notin N$ に付け替えたものをスケルトンと呼ぶ．さらに，文法 G の導出木のスケルトンを，G から生成されるスケルトンと呼ぶ．

このとき，偏導出線形文法で生成可能な w に対するあらゆる導出木のスケルトンについて，G_w はそれと同型なスケルトンとなる導出木を生成可能である．

次に，代表部分集合 Q に含まれるすべての語 w に対して G_w を構成し，以下のような文法 G_{all} を考える．

$$G_{\mathrm{all}} \;=\; \left(\bigcup_{w \in Q} R_w, \Sigma, \bigcup_{w \in Q} P_w, S_w\right)$$

ここで，S_w は任意に選択した $w \in Q$ に対する G_w の開始記号である．文法 G_{all} における非終端記号および生成規則の集合をそれぞれ，

$$\begin{aligned}
R \;&=\; \bigcup_{w \in Q} R_w \\
P_{\mathrm{all}} \;&=\; \bigcup_{w \in Q} P_w
\end{aligned}$$

と表す.

このとき，Q を代表部分集合とし，生成する言語が学習対象と等価な偏導出線形文法 $G_t = (N_t, \Sigma, P_t, S_t)$ すべてに対して，以下が成り立つ.

定理 7 以下の条件を満たすような，G_{all} の非終端記号の集合 R_{all} の部分集合 R' および R' から N_t への全射 $f : R' \to N_t$ が存在する.
(条件): P_t の任意の生成規則 $A \to uBw$ に対して，$A = f(r_1), B = f(r_2)$ かつ $r_1 \to u \cdot r_2 \cdot w$ が G_{all} の生成規則であるような $r_1, r_2 \in R'$ が存在する．ここで，$A, B \in N_t$, $u, w \in \Sigma^*$ である.
(証明) $w \in Q$ に対する G_w を構成すると，G_w から生成されるスケルトンには，G_t における w の導出木のスケルトンが含まれている．代表部分集合の定義より，Q を生成するためには G_t のすべての生成規則が使われるため，題意が成り立つ. □

したがって，偏導出線形言語に対する学習アルゴリズムは，以下の方針で構成することができる.

1. 代表部分集合 Q から線形文法 G_{all} を構成する.
2. G_{all} の非終端記号を適切に分類し，不適切な生成規則を削除する.

3.2 非終端記号の分類と不適切な生成規則の削除

本アルゴリズムでは，非終端記号の分類，および不適切な生成規則の削除を観察表 [1] を用いて行う．観察表は，$W \subseteq \Sigma^*$ と

$$\begin{aligned} R &= \{(x,y,z) \mid x,z \in \Sigma^*, y \in \Sigma^+, x \cdot y \cdot z \in Q\} \\ &\quad \cup \{(\varepsilon, w, \varepsilon) \mid w \in Q\} \end{aligned}$$

および $T : R \times \Sigma^* \to \{0,1\}$ から構成される．ここで，

$$T((u,v,w),x) = MEMBER(u \cdot x \cdot w)$$

である.

表 1 に以下の場合の観察表の例を示す.

$$\begin{aligned} L_t &= \{a^i b^i \mid i \geq 1\} \\ G_t &= (\{S, A\}, \{a, b\}, P_t, S) \\ P_t &= \{S \to Ab, A \to aAb, A \to a\} \\ Q &= \{aabb\} \\ W &= \{a, b, ab, aa, bb, aab, abb, aabb\} \end{aligned}$$

表 1: 観察表の例

R	W: a	b	ab	aa	bb	aab	abb	$aabb$
$(\varepsilon, aabb, \varepsilon)$	0	0	1	0	0	0	0	1
(a, abb, ε)	0	0	1	0	0	0	0	1
(aa, bb, ε)	0	1	0	0	0	0	1	0
(aab, b, ε)	0	0	0	0	1	0	0	0
(ε, aab, b)	0	1	0	0	0	0	0	0
(ε, aa, bb)	0	0	0	1	0	0	0	0
(ε, a, abb)	1	0	0	0	0	0	0	0
(a, ab, b)	0	0	1	0	0	0	0	1
(aa, b, b)	0	1	0	0	0	0	1	0
(a, a, bb)	1	0	0	0	0	1	0	0

観察表を利用して R を分類するために R 上の同値関係 $\overset{\pi}{\equiv}$ を定義する.

定義 8 R 上の同値関係 $\overset{\pi}{\equiv}$ を以下のように定義する．$r, r' \in R$ に対して，

$$r \overset{\pi}{\equiv} r' \iff T(r,w) = T(r',w) (\forall w \in W)$$

さらに $B(r,\pi) = \{r' \in R \mid r' \overset{\pi}{\equiv} r\}$ と定義する. □

以上の定義から，$(\varepsilon, v, \varepsilon)$ なる形の R の要素すべては，同一の $B((\varepsilon, w, \varepsilon), \pi)$ に含まれる．さらに，$W_1 \subseteq W_2 \subset \Sigma^*$ について，$\overset{\pi_1}{\equiv}$ を R, W_1, T からなる観察表における同値関係であるとし，$\overset{\pi_2}{\equiv}$ を R, W_2, T からなる観察表における同値関係であるとする．このとき，π_2 は π_1 と等しいか，より細かい.

同値関係 $\overset{\pi}{\equiv}$ を用いて G_{all} の非終端記号を分類し，その分類に合わせて生成規則も分類した線形文法 $G_\pi = (R/\pi, \Sigma, P_{\mathrm{all}}/\pi, S_{\mathrm{all}})$ を以下のように定める.

$$\begin{aligned} R/\pi &= \{B(r,\pi) \mid r \in R\}, \\ P_{\mathrm{all}}/\pi &= \{B(r_1,\pi) \to a \mid a \in \Sigma, r_1 \in R, \\ &\qquad (r_1 \to a) \in P_{\mathrm{all}}\} \\ &\cup \{B(r_1,\pi) \to aB(r_2,\pi) \mid a \in \Sigma, r_1, r_2 \\ &\qquad \in R, (r_1 \to ar_2) \in P_{\mathrm{all}}\} \\ &\cup \{B(r_1,\pi) \to B(r_2,\pi)a \mid a \in \Sigma, r_1, r_2 \\ &\qquad \in R, (r_1 \to r_2 a) \in P_{\mathrm{all}}\} \\ &\cup \{B(r_1,\pi) \to aB(r_2,\pi)b \mid a, b \in \Sigma, r_1, \end{aligned}$$

$$r_2 \in R, (r_1 \to a r_2 b) \in P_{\mathrm{all}}\}$$

$$S_{\mathrm{all}} \quad = \quad B((\varepsilon, w, \varepsilon), \pi)$$

ここで $w \in Q$ は Q から任意に選択したある終端記号列である.

線形文法 G_π から以下の操作により生成規則を削除する. ここで, $A, B \in R/\pi$ であり, $a \in \Sigma$ である.

操作 9 $u_1, u_2 \in \Sigma^*$, $A, B \in R/\pi$ について, $A \to u_1 B u_2$ は P_{all}/π に含まれている生成規則であるとし, $r_A \in B(r_A, \pi) = A$, $r_B \in B(r_B, \pi) = B$ であるとする. このとき, $T(r_A, u_1 w u_2) \neq T(r_B, w)$ なる $w \in W$ が存在するならば, $A \to u_1 B u_2$ を P_{all}/π から取り除く. □

この操作は, B が w を生成できるのに対して, A が $u_1 w u_2$ を生成でいない場合, もしくはその逆の場合に $A \to u_1 B u_2$ なる規則を不適切と見なしている. さらに, 以下の操作を行う.

操作 10 $u_1, u_2 \in \Sigma^*$, $A, B \in R/\pi$ について, $A \to u_1 B u_2$ は P_{all}/π に含まれている生成規則であるとし, $r_A \in B(r_A, \pi) = A$, $r_B \in B(r_B, \pi) = B$ であるとする. このとき, $T(r_B, w) = 1$ であっても, $w \notin L_{G_\pi}(B)$ であるような $w \in W$ が存在するならば, $A \to u_1 B u_2$ を P_{all}/π から取り除く. □

この操作は, B が w を生成すべきところに, 生成規則が十分そろっていないという場合, その B を右手側に含む生成規則を削除する. この操作を行って削除される生成規則が存在する場合, 新たに削除の条件を満たす生成規則が発生する場合がある. そこで, 本操作を $|P_{\mathrm{all}}/\pi|$ 回繰り返すことにより, 新たにこの条件を満たす生成規則を取り除くことができる.

上記の操作を $|P_{\mathrm{all}}/\pi|$ 回行った後の P_{all}/π を規約であると呼ぶ. また, 既約な生成規則の集合 P_{all}/π に含まれる $A \to a$ $(a \in \Sigma)$ なる形の生成規則すべての集合を P_Σ と表す. このとき, 以下の補題が成り立つ.

補題 11 $P_\Sigma \subseteq P_1 \subseteq P_{\mathrm{all}}/\pi$ を満たす生成規則の集合 P_1 および, ある $y \in Q$ について,

$$G_1 = (R/\pi, \Sigma, P_1, B((\varepsilon, y, \varepsilon), \pi))$$

なる線形文法 G_1 は偏導出線形文法であるとする. このとき, G_1 において到達可能かつ live である $B(r, \pi) \in R/\pi$ および $w \in W$ に対して,

$$T(r, w) = 1 \iff B(r, \pi) \overset{*}{\underset{G_1}{\Rightarrow}} w$$

が成り立つ.

(証明) w の長さに関する帰納法で証明する. $|w| = 1$ の場合, P_1 の仮定 $P_\Sigma \subseteq P_1$ より題意を満たす. 任意の $|w| \leq n$ で題意が成り立つとし, いま $|w| = n+1$ であるとする. P_{all}/π は既約であるので, P_1 も既約である. $T(r, w) = 1$ ならば, P_1 は既約であることより, ある生成規則 $B(r, \pi) \to u_1 B(r', \pi) u_3$ について $T(r', w') = 1$ である. ここで, $w' \in \Sigma^+$ は $u_1 w' u_3 = w$ なる記号列であり, $|u_1| + |u_3| \geq 1$ である. 帰納法の仮定より, $T(r', w') = 1$ であり, かつそのときに限り $w' \in L_{G_1}(B(r', \pi))$ である. すなわち, $T(r, w) = 1$ であり, かつそのときに限り $w \in L_{G_1}(B(r, \pi))$ であり, 題意が成り立つ. □

上記補題より, $P_\Sigma \subseteq P_1 \subseteq P_{\mathrm{all}}/\pi$ であるような任意の生成規則の集合 P_1 から構成された偏導出線形文法は, 観察表の結果と矛盾しない. すなわち, そのような偏導出線形文法をすべて数え上げれば, その中に学習対象と等価な文法が存在する. しかし, 一般にそのような数え上げを多項式時間で行うことは難しい.

そこで, 基底文法族と呼ばれる偏導出線形文法の集合 $\mathbf{G}$ を以下の手順で構成する.

1. $P_0 = P_\Sigma$ とする.
2. すべての $A \in R/\pi$ および, $a = b$ も含めた 2 種類の終端記号のすべての組み合わせ $a, b \in \Sigma$ に対して, $A \to aBb$ なる形の生成規則を P_{all}/π から任意に 1 つずつ選び, 選び出した生成規則をすべて P_0 に加える.
3. P_{all}/π に含まれる $A \to aB$, $A \to Bc$ なる形の生成規則すべてについて, それを P_0 に加えても偏導出線形文法の定義を満たしている場合には, P_0 に加える. ここで, 加える順番は任意である.
4. 上記手順で構成した生成規則の集合を用いて, 偏導出線形文法 $G_0 = (R/\pi, \Sigma, P_0, S_{\mathrm{all}})$ を定める.
5. P_{all}/π に含まれる生成規則 $A \to u_1 B u_2$ $(u_1, u_2 \in \Sigma^*)$ について, $P(A \to u_1 B u_2)$ を次

のように定める．P_0' を P_0 に $A \to u_1Bu_2$ を加えた生成規則の集合とする．$P(A \to u_1Bu_2)$ は，P_0' が偏導出線形文法となるように $A \to u_1Bu_2$ 以外の生成規則を順次削除したものとする．この削除は，どのような順序で行っても同じ結果となる．

6. $\mathbf{G} = \{G(A \to u_1Bu_2) \mid G(A \to u_1Bu_2) = (R/\pi, \Sigma, P(A \to u_1Bu_2), S_{\mathrm{all}}), (A \to u_1Bu_2) \in P_{\mathrm{all}}/\pi\}$ とする．

基底文法族 $\mathbf{G}$ に対して，学習者は以下の処理を行う．

- すべての $A \in R/\pi$ および $G_1 \neq G_2$ であるすべての $G_1, G_2 \in \mathbf{G}$ に対して，

$$L_{G_1}(A) = L_{G_2}(A)$$

であるか否かを判定する．

- 前ステップのすべての等価性判定が等価ならば，任意の $G \in \mathbf{G}$ を提示し，そうでなければ，非等価の証拠となる記号列 $w \in (L_{G_1}(A) - L_{G_2}(A)) \cup (L_{G_2}(A) - L_{G_1}(A))$ のすべての部分記号列を観察表の W に加えて，観察表を再構成する．

偏導出文法の効率的な等価性判定可能性については，未解決問題である．全体の学習アルゴリズムは，図 1 となる．

```
INPUT : L_t の代表部分集合 Q
OUTPUT : 仮説文法 G_h
begin
  R := {(x,y,z) | x,z ∈ Σ*, y ∈ Σ+, x·y·z ∈ Q};
  W := {y ∈ Σ+ | x,z ∈ Σ*, x·y·z ∈ Q};
  do
    観察表を構成し，同値関係 ≡_π を定める;
    操作 9, 10 により不要な規則取り除き，
    P_all/π を既約にする;
    基底文法族 G を求める;
    W' := ∅;
    for every pair of G_1, G_2 ∈ G and
    every A ∈ R/π do
      w ∈ (L_G1(A) − L_G2(A)) ∪ (L_G2(A) − L_G1(A))
      かつ |w| が多項式長である w を求める;
      W' := W' ∪ {w};
    done
    for all w ∈ W' do
      W := W ∪ {y ∈ Σ+ | x,z ∈ Σ*, x·y·z = w};
    done
  while (W' ≠ ∅);
  任意の G ∈ G を出力する;
end.
```

図 1: 学習アルゴリズム

3.3 アルゴリズムの正当性と停止性

以上のアルゴリズムにおいて，以下の補題が成り立つ．

補題 12 任意の $A \in R/\pi$ および，任意の $G_1, G_2 \in \mathbf{G}$ $(G_1 \neq G_2)$ について，$L_{G_1}(A) = L_{G_2}(A)$ であるとき，いかなる $G \in \mathbf{G}$ についても $L(G) = L_t$ が成り立つ．

(証明) 代表部分集合の定義より，いかなる $A \in N_t$ に対しても，ある $B((u_1,u_2,u_3),\pi) \in R/\pi$ が存在し，

$$S_t \overset{*}{\underset{G_t}{\Rightarrow}} u_1Au_2$$

ならば，かつそのときに限り

$$S_{\mathrm{all}} \overset{*}{\underset{G}{\Rightarrow}} u_1B((u_1,u_2,u_3),\pi)u_2$$

が成り立つ．このような $(u_1,u_2,u_3) \in R/\pi$ を r_A と表す．以後，$w \in \Sigma^*$ について，$|w|$ に関する帰納法を用いて

$$w \in L_G(B(r_A,\pi)) \iff w \in L_{G_t}(A)$$

を示す．

$|w| = 1$ の場合，$G \in \mathbf{G}$ の生成規則の集合はすべて P_Σ を含んでいるので，題意を満たす．次に $|w| = n$ の場合に題意が成り立つと仮定し，いま $|w| = n+1$ であるとする．

生成規則の集合 P_t のいかなる生成規則 $A \to w_1Bw_2$ $(A,B \in N_t,\ w_1,w_2 \in \Sigma^*)$ に対しても $B(r_A,\pi) \to w_1B(r_B,\pi)w_2$ なる生成規則が P_{all}/π に含まれている．したがって，任意の $A \in N_t$ および $w \in \Sigma^+$ $(|w| = n+1)$ について，ある $G' \in \mathbf{G}$ が存在し，

$$w \in L_{G_t}(A) \iff w \in L_{G'}(B(r_A,\pi))$$

である．仮定より，任意の $G \in \mathbf{G}$ について，

$$w \in L_G(B(r_A,\pi)) \iff w \in L_{G'}(B(r_A,\pi))$$

であるので，A を S_t と置き換えれば題意が満たされる． □

一方，$L_{G_1}(A) \neq L_{G_2}(A)$ であるような $G_1, G_2 \in \mathbf{G}$ および $A \in R/\pi$ が存在するならば，以下の補題が成り立つ．

補題 13 R, W, T を観察表とし，そこから得られた生成規則の集合を P_{all}/π とする．ある $G_1, G_2 \in \mathbf{G}$ および $A \in R/\pi$ について，$w \in (L_{G_1}(A) - L_{G_2}(A)) \cup (L_{G_2}(A) - L_{G_1}(A))$ であるとする．

いま，w により観察表が更新され，$R, W_w = W \cup \{w' \in \Sigma^* \mid uw'v = w, u, v \in \Sigma^*\}, T$ からなるとする．更新された観察表における同値関係を $\overset{\pi'}{=}$ としたとき，以下のいずれかが成り立つ．

1. 以下を満たすような $u, v \in \Sigma^*$ および $r_0, r_1 \in R$ が存在する．
   - $B(r_0, \pi) \rightarrow uB(r_1, \pi)v$ なる生成規則が P_{all}/π に含まれるが，
   - $B(r_0, \pi') \rightarrow aB(r_1, \pi')v$ なる生成規則は P_{all}/π' に含まれない．
2. 新しい観察表における同値類の集合 π' は，π より細かい．

(証明) 題意が満たされないとすると，π と π' は等しく，$P_{\mathrm{all}}/\pi = P_{\mathrm{all}}/\pi'$ である．したがって，任意の $r \in R$ について $L_{G_1}(B(r, \pi)) = L_{G_1}(B(r, \pi'))$ かつ $L_{G_2}(B(r, \pi)) = L_{G_2}(B(r, \pi'))$ である．ところが，仮定より一般性を失わずに $w \in L_{G_1}(B(r, \pi')) - L_{G_2}(B(r, \pi'))$ であるとすると，新しい観察表においては，補題 11 より $T(r, w) = 1$ かつ $T(r, w) = 0$ である．これは矛盾であり，したがって題意を得る．□

補題 13 より，図 1 における学習アルゴリズムは，高々 $|P_{\mathrm{all}}|$ 回の繰り返しの後，$W' = \emptyset$ となる．したがって補題 12 より，学習アルゴリズムは正しい仮説を出力して停止する．このときの時間計算量は，偏導出線形文法の等価性判定に必要な計算量の多項式倍である．

定理 14 任意の偏導出線形言語は，所属性質問と代表部分集合から厳密学習可能である．ここで，学習に要する計算量は，偏導出線形文法の等価性判定問題に要する計算量に関する多項式で抑えることができる．□

4 まとめ

本研究において，偏導出線形言語に対して，所属性質問と代表部分集合を用いた厳密学習アルゴリズムを示した．その時間計算量は，それぞれの言語を表す文法族における等価性判定問題の時間計算量に関する多項式となっている．

本アルゴリズムは観察表を用いて非終端記号を分類し，学習対象を表せる任意の文法に関する多項式サイズの基底文法族が構成できることが効率的な学習可能性につながっている．今後の課題として，上記のような条件を満たした文脈自由文法の部分族の発見，および条件のさらなる一般化などが挙げられる．

また，サンプリングから代表部分集合を確率的に構成する手法による PAC 学習との関連性も今後の課題として挙げられる．

参考文献

[1] D. Angluin, Learning regular languages from queries and counterexamples, Inf. & Comp. **75** (1987) 87–106.

[2] M. A. Harrison, Introduction to Formal Language Theory, Addison-Wesley, Reading, MA, 1978.

[3] J. E. Hopcroft, J. D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison-Wesley, Reading, MA, 1979.

[4] C. de la Higuera, J. Oncina, Inferring deterministic linear languages, 15th Ann. Conf. on Computational Learning Theory - COLT 2002, LNAI **2375** (2002) 185–200.

[5] Y. Takada, A hierarchy of language families learnable by regular language learning, Inf. & Comp. **123** (1995) 138–145.

[6] Y. Tajima, E. Tomita, A polynomial time learning algorithm of simple deterministic languages via membership queries and a representative sample, Proc. 5th Int. Coll. on Grammatical Inference - ICGI 2000 : LNAI **1891** (2000) 284–297.